

## JRC TECHNICAL REPORTS

# Collection and filtering of relevant knowledge

*The Knowledge Browser  
Feasibility study*

BRYTON DIAS MARQUES, FSBDM

2016

This publication is a Technical report by the Joint Research Centre (JRC), the European Commission's science and knowledge service. It aims to provide evidence-based scientific support to the European policymaking process. The scientific output expressed does not imply a policy position of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use that might be made of this publication.

**Contact information**

Name: Fernando Sergio BRYTON DIAS MARQUES

Address: IPSC – Institute for the Protection and Security of the Citizen

Email: fernando.bryton@jrc.ec.europa.eu

Tel.: +39 033 278 6604

**JRC Science Hub**

<https://ec.europa.eu/jrc>

JRC105006

EUR 28375 EN

PDF	ISBN 978-92-79-64800-7	ISSN 1831-9424	doi:10.2788/33172
Print	ISBN 978-92-79-64799-4	ISSN 1018-5593	doi:10.2788/180063

Luxembourg: Publications Office of the European Union, 2016

© European Union, 2016

The reuse of the document is authorised, provided the source is acknowledged and the original meaning or message of the texts are not distorted. The European Commission shall not be held liable for any consequences stemming from the reuse.

How to cite this report: Bryton, F., *Collection and filtering of relevant knowledge*, EUR 28375 EN, doi: 10.2788/33172

All images © European Union 2016

## Contents

Abstract .....	1
1. Introduction .....	5
2. Current approach .....	7
2.1 The "Gather and share" process .....	8
2.2 The "Retrieve" process .....	9
2.3 Solution requirements .....	11
3. Enterprise Search solutions .....	11
3.1 Background .....	11
3.2 Core concepts .....	12
3.3 Typical features .....	14
3.4 Expected benefits .....	14
3.5 Challenges .....	16
4. Proposed approach .....	17
4.1 Development from scratch .....	18
4.2 Reuse of functionally relevant OSS components .....	19
4.3 Reuse of functionally relevant systems .....	19
4.4 Crosscutting issues .....	20
4.5 Proposal .....	20
5. Evaluation .....	21
5.1 Evaluation of information retrieval systems .....	21
5.2 Quantitative evaluation of the proposed solution .....	22
5.2.1 Scenario 1: the proposed solution is not developed .....	23
5.2.2 Scenario 2: the proposed solution is developed and adopted .....	25
5.2.3 Comparative analysis .....	26
Conservative perspective .....	26
Optimistic perspective .....	27
5.3 Qualitative evaluation of the proposed solution .....	28
6. Conclusions and recommendations .....	28
7. References .....	31
List of abbreviations and definitions .....	33
List of figures .....	35
List of tables .....	37



## Abstract

Improving the way in which data, information and knowledge are gathered and shared is a priority for the European Commission (EC), the Joint Research Centre (JRC) and the Knowledge Centre for Migration and Demography (KCMD). This is expected to increase productivity, foster teamwork and to contribute to overcome silo mentalities and harness synergies between portfolios, hence to enable the Commission to come up with fast and effective solutions for the challenges it faces.

The Knowledge Browser (KB) herein proposed is a technological solution which aims to enable users from the JRC, the Commission and from other organizations to search across a selected collection of internal and external sources of data, information and knowledge, relevant for policy making, in a way that is more effective, efficient and independent than traditional approaches. For that purpose, the KB will build on and promote the usage of existing tools such as the KCMD Data Catalogue and the Europe Media Monitor.

As such, the KB solution addresses directly the objectives and actions foreseen by the Commission's recent communication on Data, Information and Knowledge Management. Moreover, the KB is envisioned, in the first place, to overcome challenges identified by the KCMD. From a JRC point of view, the primary JRC "priority nexus" it is addressing is the one on "Migration and Territorial Development". However, considering its cross-cutting nature, the KB has the potential to contribute to the other priority nexuses as well. The KB also contributes directly to two of the pillars of JRC's scientific excellence: sharing and productivity.

This document analyses the feasibility of the KB. To do so, the departure point is an analysis of some of the present KCMD processes, which highlights their main challenges and supports the definition of the KB general requirements. Then, the best approach to develop the KB is first selected, among different alternatives, and evaluated. This evaluation, which is both quantitative and qualitative, includes a cost-benefit analysis based on an estimation of the value of the time that will be saved by introducing the KB in the present KCMD processes.

The KB is a strategic initiative expected to bring benefits over the short and long term, to the KCMD, the JRC and the Commission, which largely exceed its costs. It has been estimated that the KB may save work to the KCMD up to the equivalent of 3 person.year and 150.000 per year. However, considering the number of knowledge management and production units of the JRC and assuming they face similar challenges, the KB could bring savings up to 7.5 million euros per year.

On the other hand, the qualitative benefits include the enhanced visibility and relevance of the JRC, the KCMD and the Commission, the enhanced performance of the KCMD, the JRC and of its partners (inside and outside the Commission), and the increased motivation of researchers and other knowledge workers.

The development of a prototype of the KB, based on open source technology and implementing relevant requirements, is therefore recommended, for the duration of six months and provided that the necessary material and human resources are made available in useful time. The authors are aware that at the moment of this writing similar JRC-level work is being undertaken to improve the way and the tools through which data, information and knowledge are gathered and shared in the EC. This work is intended to complement that wider parallel experience by focusing on one knowledge management solution for the KCMD.



*"The most important contribution management needs to make in the 21st century is to increase the productivity of knowledge work and knowledge workers"*

(Peter F. Drucker, 1999)





## 1. Introduction

Improving the way in which data, information and knowledge are gathered, managed, shared and preserved is inherent to the necessary modernisation of the Commission's ways of working. This is expected to foster teamwork and to contribute to overcome silo mentalities and harness synergies between portfolios, hence to enable the Commission to come up with fast and effective solutions for the challenges it faces (European Commission, 2016a).

This demands for simultaneous actions in three different planes. The first one is that of effective sharing and exploitation of the needed data, information and knowledge; the second one is that of removing the barriers to working together; and the third one is that of developing methods and tools to support these new working habits and organisational culture (European Commission, 2016a).

In this context, among the main areas for improvement that have been identified, we highlight the following (European Commission, 2016a):

- Information retrieval and delivery;
- Working together and sharing information and knowledge.

On one hand, improving information retrieval and delivery is needed given the present difficulty to find and retrieve the necessary information which, together with the likely existence of duplication and inconsistencies, reduces productivity. Therefore, data and information should be made searchable, easily retrievable and as widely available as possible across the organization. For these reasons, one of the actions defined is to (European Commission, 2016a):

- Develop the capability to search easily across different systems in order to find and retrieve all the relevant information held by the Commission, irrespective of where that information is stored or the underlying technology.

On the other hand, one of the defined steps for working together and sharing information and knowledge is the setup of specific knowledge and competence centres for issues that fall under the policy priorities of the Commission, including thematic areas of the European Semester. Consequently, the following action has been defined (European Commission, 2016a):

- The JRC initiative to develop Knowledge Centres and Competence Centres for priority policy areas will be further developed.

In this context, the mission of the JRC Knowledge Centres is to inform policy makers about the status and findings of the latest scientific evidence, legitimate disagreements in the scientific community and about scientific limits and uncertainties. It is also to map and, when appropriate, fill in knowledge gaps; as well as coordinate the supply of knowledge by consolidating knowledge from across the scientific community, practitioners, policy makers and other relevant stakeholders. All of this in a context where its resources will not grow, which demands for an increase of its efficiency (Joint Research Centre, 2016).

In particular, the JRC Knowledge Centre for Migration and Demography (KCMD), in order to enhance the knowledge base and support the work on migration of the relevant Commission services, EU Member States and their strategic partners, shall develop analytical and networking activities accompanied by a repository of relevant research projects and new initiatives to deepen knowledge. By doing so, the KCMD will add value by providing easy and quick access to existing relevant, structured and validated knowledge and activities on migration and demography inside the EU (JRC Task Force on Migration, 2016).

The KCMD shall also provide a platform where the needs and questions of Commission services and EU Member States are served through a robust and relevant evidence base,

research and studies, facilitated and powered through strong networking across Europe and internationally (JRC Task Force on Migration, 2016).

Accordingly, the KCMD, during its first two years of operation, will focus on three types of activities, among which are those to which the initiative herein proposed aims to contribute (JRC Task Force on Migration, 2016):

- Building the evidence base, conducting analysis and foresight;
- Exploiting the knowledge base and facilitating uptake by stakeholders.

On one hand, to build the evidence base and conduct analysis and foresight, the KCMD shall, among other, bring together and make sense of fragmented data and information inside and outside of the Commission. On the other hand, the KCMD website shall be a central platform for sharing the knowledge of the commission and its partners. In particular, the portal shall facilitate the transfer and usage of the results of research projects (JRC Task Force on Migration, 2016).

However, gathering the relevant data, information and knowledge for these purposes is a very demanding and time consuming initiative. Firstly, because such resources are scattered across multiple organizations worldwide. Secondly, because they are made available in multiple different formats and infrastructures. Thirdly, because new organizations and new resources are frequently arising and, last but not least, because a minimum of domain expertise is required to do it properly.

Given the present magnitude and impact of the human migration phenomena worldwide, multiple organizations (within the Commission as well) are presently making available data, information and knowledge on this topic. As an example, from a small pilot project conducted on the climate change and migration nexus, it was possible to identify 6 commission services and 24 other organizations in these conditions. It is worth noting that this nexus is a subtopic of migration, and also that no local or national organizations have been considered, which would have highly increased these numbers.

These organizations make available the data, information and knowledge in many different formats such as the ones identified during the abovementioned pilot project - databases, reports and news - seldom complying with applicable standards, if existing. Moreover, each organization makes such artefacts available in different locations within their own websites, which varies according to their information management policies and procedures. This adds an additional layer of complexity that means that the access to the artefacts implies knowing where to collect them in each particular site. These aspects make it very hard to identify the relevant artefacts.

Even though it becomes possible to gather, in a reasonable amount of time, the relevant artefacts, it must be considered that new organisations emerge constantly, and that both the existing and the new ones will continue to produce new data, information and knowledge. This implies three things. First, that any collection of such artefacts will be rapidly dated; second that it will be hard to identify the new artefacts among the previously identified ones and, finally, that these challenges tend to increase with time.

Last, but not least, a minimum level of domain expertise is required for gathering the relevant data, information and knowledge. In order to properly find and select the relevant artefacts, it is at least necessary to have a basic understanding of the concepts involved.

Consequently, it is urgent to introduce automation into the processes of the KCMD which comprise gathering and sharing data, information and knowledge, without which their efficiency, effectiveness and quality will be severely hampered in the short term.

With this in mind, the Knowledge Browser is the envisioned technological solution for collecting and providing migration data, information and knowledge relevant for EU and MS policy and decision makers, which will build on and interact with internal (Commission) and external (strategic partners and other stakeholders) existing databases and web sites.

Its purpose is to support the KCMD in gathering and sharing the relevant migration and demography knowledge for sound policy making; that is, moving the right knowledge, to the right people at the right time (O'Dell & Grayson, 1998). Namely, with this tool, its users will be able to, more quickly and independently, retrieve the relevant migration artefacts, whether these reside in or out of the organization.

As such, the KB may contribute to the following actions (European Commission, 2016b):

- Action 1.1 : Develop capability to search easily across different systems;
- Action 2.A.1 : Pilot project on cross-DG collaborative policy making.

In this context, this feasibility study has presently been carried out, with the purpose of identifying a good approach for implementing it, understanding its cost-effectiveness as well as other aspects of its implementation. Afterwards, the intention is to design and prototype the KB, primarily making it available for the KCMD, and then considering the possibility of extending it to support equivalent knowledge management processes in other areas of the JRC.

The rest of this study is organized as follows. In section 2, the state-of-the-art regarding how migration data, information and knowledge are presently gathered and shared by the KCMD is presented and discussed, highlighting particularly the existing challenges which the KB can help to address and defining the KB general requirements. Section 3 introduces the state-of-the-art of enterprise search and related technologies. Then, in section 4, the most relevant implementation alternatives are discussed and the best one is selected, which will be evaluated in section 5, from the qualitative and quantitative perspectives. This study is then finished in section 7 by presenting the conclusions and recommendations.

## 2. Current approach

There are two key KCMD processes relevant for the matter at hand. These are the "Gather and share" (data, information and knowledge artefacts) process and the "Retrieve" (those artefacts) process. The purpose of this section is to describe these processes, in broad terms, and point out their challenges, which will be used to identify the opportunities for the KB to introduce improvements, hence the foundations for its requirements.

Among the different tools that are already used within these processes, four are worth being mentioned. These are the KCMD website<sup>1</sup>, the data catalogue<sup>2</sup>, the information catalogue<sup>3</sup> and the European Media Monitor (EMM) NewsExplorer<sup>4</sup>.

The KCMD website is the entry point for obtaining migration and demography data, information and knowledge relevant for policy making in these domains. Additionally, other relevant information can be found in this site, such as the strategy of the KCMD, knowledge gaps, analysis and foresight, and the KCMD partnerships.

The data catalogue is a table of selected data sources relevant to Migration and Demography policies. Each data source is listed with its summary description, the link to its web site and other metadata. The catalogue will include official EU and international statistics, as well as important data sets at Member State level.

The information catalogue is a list of all knowledge and information sources considered relevant for migration and demography policy making. Each of the entries in this list is accompanied by, at least, its title, a short description, the author identification, the date in which it was created, a link pointing to the original document and a set of keywords

---

<sup>1</sup> <https://ec.europa.eu/jrc/en/migration-and-demography>

<sup>2</sup> <http://bluehub-ckan-dev.jrc.it:5000/>

<sup>3</sup> <https://ec.europa.eu/jrc/en/migration-and-demography/knowledge/information>

<sup>4</sup> <http://emm.newsexplorer.eu>

which helps in characterising it. The information catalogue entails many different types of artefacts, such as reports, infographics and news.

The NewsExplorer automatically generates daily news summaries, allowing users to see:

- The major news stories (news clusters) in various languages for any specific day and to compare how the same events have been reported in the media written in different languages;
- The list of most mentioned names and find further automatically derived information (e.g. variant name spellings, titles and phrases, list of the most recent articles and list of related persons and organisations).

## **2.1 The “Gather and share” process**

The “Gather and share” process, depicted in Figure 1, is the process by which the KCMD staff will make available the migration and demography data, information and knowledge that is relevant for policy making. Within this process, the following activities are presently performed:

- 1) Gather stakeholders: consists in performing a search through the available means (i.a. internet, bibliography) with the purpose of identifying the most relevant stakeholders (e.g. research centres, practitioners, policymakers, news media) of a particular topic within the wider migration and demography domain. A common characteristic to these stakeholders is that they are producers of data, information or knowledge relevant for policy making. The output of this activity is a list of those stakeholders composed by some metadata (e.g. name, description, type, link to website);
- 2) Gather sources (artefacts): consists in going thoroughly through the websites of the previously identified stakeholders (with the help of EMM in the case of news sources) with the purpose of identifying relevant sources (e.g. databases, reports, infographics, and news’ articles) of data, migration and knowledge. The output of this activity is a list of those sources composed by basic metadata (e.g. name, type, link to artefact);
- 3) Enhance sources (artefacts): consists in overviewing each of the previously identified potential artefacts and increasing their metadata (e.g. description, author, keywords), with the objectives of adding relevant and useful metadata to the artefacts and further verifying their relevance. Consequently, some of the artefacts may be removed from the list. The output of this activity is, therefore, the original list of sources enhanced;
- 4) Share sources (artefacts): consists in making available, in a comprehensive way, the metadata of the sources previously gathered, which will be used by the “customers” to find and decide on the relevance of the available sources to their specific task, and also to retrieve those from their original repositories. The data artefacts (databases) are made available in the data catalogue, and the information sources are made available in the information catalogue abovementioned. The outputs of this task are the catalogues (data and information) up to date.



**Figure 1: The “Gather and share” process**

The success of this process could be measured by the ratio between the quantity of relevant artefacts catalogued and the total quantity of artefacts catalogued. This implies that the more artefacts are catalogued, the higher is the probability of cataloguing the relevant ones, and also implies that the less demanding is the cataloguing activity, in terms of selection, the highest is the probability of cataloguing irrelevant artefacts.

In simple terms, this process demands time and expertise. Time to gather plenty of artefacts and to analyse them, and sufficient expertise to effectively separate the relevant ones from the remainder. Additionally, to minimize human errors and subjectivity this process could benefit from a high degree of automation.

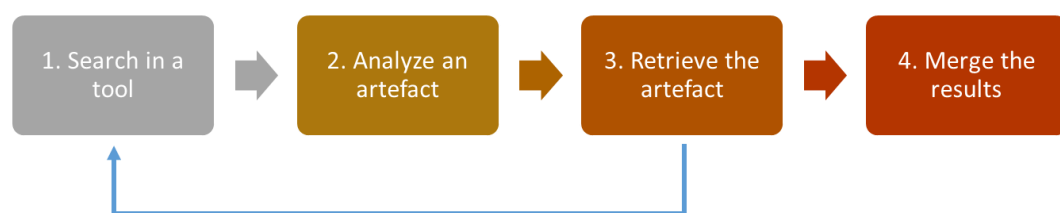
As such, the biggest challenge this process faces comes from the amount of artefacts available versus the number of available resources to carry out the activities it entails. The fact that the resources to carry out those activities are limited and will not increase in the near future is assumed from start. Therefore, the issue comes from the vast number of stakeholders, websites and artefacts available, which increases with time. Consequently, each of the activities previously identified, faces the following specific challenges and limitations:

- 1) Gather stakeholders: When a new topic is dealt with, for the first time, identifying the relevant stakeholders can be a time consuming and subjective task. On the other hand, when the topic is not new, the difficulty lies on identifying new ones;
- 2) Gather sources (artefacts): In the beginning, gathering the relevant artefacts is a very time consuming task. However, with time, this becomes worse with the additional task of identifying what is new. Additionally, since this implies selection and cataloguing activities, it is also vulnerable to human error and subjectivity;
- 3) Enhance sources (artefacts): Likewise, enhancing the artefacts is also a very time consuming task, which is also vulnerable to human error and subjectivity, since it also entails decisions regarding how to classify and which artefacts to keep. Additionally to enhancing the metadata of the new artefacts, it is necessary to verify the metadata (at least partially) of all the artefacts previously catalogued, because the catalogues cannot be dated, or even worse, pointing to nowhere. This can be the case where links have become broken, for example because documents have been relocated or simply deleted;
- 4) Share sources (artefacts): This is also a very time consuming task, which is mostly vulnerable to human error, introducible when making the metadata available in the specific platforms.

## **2.2 The “Retrieve” process**

The “Retrieve” process, depicted in Figure 2, is the process by which the “customers” of the KCMD (inside and outside the Commission) presently find and access the migration and demography data, information and knowledge that is relevant for supporting policy making. This is done both by using the KCMD aforementioned tools and also other tools, which are internal (e.g. Connected, Pubsy, CORDIS) and external (e.g. Google, Scopus) to the Commission. Therefore, this process entails the following activities:

- 1) Search in a tool: consists in accessing and browsing through a selected tool searching for artefacts of interest. The output of this activity is a list of potential artefacts that meet the search criteria;
- 2) Analyse an artefact: consists in browsing through the metadata available for a particular artefact, with the purpose of deciding on whether to retrieve it or not;
- 3) Retrieve the artefact: consists in browsing to the place (e.g. website) where the artefact is made available, with the purpose of retrieving it;
- 4) Merge the results: Consists in putting together all the artefacts retrieved by using different tools and, at least, eliminating duplications.



**Figure 2: The “Retrieve” process**

In this context, the biggest challenge is that each user has to use many different tools to retrieve the relevant data, information and knowledge. Consequently, the user needs to be familiarized with each of the relevant features offered by each tool which are, most likely, different. Additionally, after looking through all tools selected, the user will still have to put all the results together and, at least, eliminate duplications.

The process is as much successful as the resulting artefacts are found relevant and retrieved by the user. In other words, the result of a search should present, as much as possible, the artefacts which are most relevant for the user, according to the criteria established, instead of as many artefacts as possible.

The success of this process is also inversely proportional to the time taken to find the relevant artefacts. Aspects that can influence this are, for example, the usability of the features offered and the way the information is organized and displayed by each tool. Naturally, the more tools are used, the more time is taken to obtain the relevant artefacts.

Consequently, the challenges faced in each of the activities previously identified, can be summarized as follows:

- 1) Search in a tool: the features and overall usability of each tool have to be well known by the user to explore them in an efficient and effective way;
- 2) Analyse an artefact: the information available from the different artefacts may be presented in different ways, depending on the tool used;
- 3) Retrieve the artefact: the links to the artefact may be broken, hence requiring to look for the same artefact in a different place;
- 4) Merge the results: putting all the results together is time consuming and proportional to the number of tools used. Additionally, finding duplicates among the results produced by different tools can also be very time consuming and may lead to errors.

## **2.3 Solution requirements**

The KB aims to be an automatic tool for collecting and providing data, information and knowledge relevant for migration and demography policy support, which will build on and interact with internal (Commission) and external (strategic partners and other stakeholders) existing databases and web sites, with the purpose of supporting the KCMD to overcome or mitigate some of its present challenges. Therefore, and considering the above, its high-level functional requirements are:

1. The KB shall enable the user to specify a search criteria and retrieve the list of corresponding results, obtained from different data sources, by using a single user interface;
2. The KB shall provide, in the results list, as much metadata as possible to enable the user to decide whether a certain artefact is relevant or not without having to leave the KB;
3. The KB shall be able to search both within Commission and external data sources;
4. The KB shall enable the user to specify which data sources should be used within the search;
5. The KB shall be capable of searching not only within different information systems and websites, but also within the files they contain;
6. The KB shall enable the user to filter the results obtained according to the metadata available;
7. The KB shall enable the user to retrieve the artefacts of interest from the original data source;
8. The access control to the artefacts mentioned in the results of a search is not responsibility of the KB but of the specific original data sources, as needed;
9. The KB shall not present duplicate results, unless that is explicitly required by the user;
10. The KB shall not present results which cannot be retrieved by the user (e.g. broken links);
11. The KB shall enable the user to subscribe for a periodic list of results which correspond to a specific search criteria;
12. The KB shall enable the user to export the results list to a file (e.g. CSV);
13. The KB could keep track of the users and their interests, and use this information to enhance its performance;
14. The KB could enable the user to authenticate for different data sources, hence making possible to widen the search to restricted access contents and data sources.

## **3. Enterprise Search solutions**

### **3.1 Background**

According to a study performed almost 20 years ago (Crabtree, Fox, & Baid, 1997), the percentage of time spent in information gathering was between 12% (12.2%) and 14% (13.7%). The same study also concluded that information gathering was rated the most frustrating activity (3.33 in a scale from 1 - most frustrating to 7 - least frustrating) among other activities considered such as negotiation, documentation, support/consulting, planning and problem solving/thinking.



A more recent study refers that the time spent searching for information averages 8.8 hours per week and analysing it consumed an additional 8.1 hours. (White & Nikolov, 2013). Another study (Dourado, 2016) refers that knowledge workers, that is workers whose main capital is knowledge (Davenport, 2005), are estimated to spend between 6.5 and 8.8 hours per week searching for information. Yet another study refers that Knowledge workers in enterprises spend on average 5 hours per week, or 12% of their time, searching for content. (Bughin et al., 2011). These studies situate the time spent in information gathering between 12% (5h) and 22% (8.8h).

The time spent searching for information also depends on the number of systems used for that purpose. On this, a study reports that 57% of all researchers stated that they needed access to four or more systems to perform their work and only 45% of all researchers rated their current process of finding information as being somewhat or very beneficial. (Solinger & Schubmehl, 2016).

The amount of information to be sought also influences the time taken in searching it. By July of 2008, more than one trillion unique URLs were indexed by Google, the number having grown by 44% annually during the preceding ten years (Bughin et al., 2011). According to research in the USA, large companies (i.e. with more than 1,000 employees) have accumulated over 100 terabytes of information, and any have more than 1 petabyte. Right now, nearly all information is born (or immediately turned) digital; the amount of data is expected to reach 35 Zettabytes in 10 years. (White & Nikolov, 2013). A different study reports that the amount of digital information will grow by a factor of 44 annually from 2009 to 2020 (Bughin et al., 2011).

Either way, the quantity of information involved is enormous. But even more worrying is its structure: it is estimated that about 80% of the information stored is either unstructured or has no adequate metadata for the needs of employees (White & Nikolov, 2013). To further aggravate this, enterprise repositories of unstructured information are growing rapidly because of the widespread adoption of social media, increased compliance and regulatory requirements and a lack of resources to remove redundant information (White & Nikolov, 2013).

Surveys indicate that senior managers are aware of the importance of unstructured information but few are taking action to provide employees with adequate tools to access this information. (White & Nikolov, 2013).

In fact, in spite nearly three quarters of organizations say that enterprise search is vital to their productivity, effectiveness and compliance, only 11 percent are using it (Formtek, 2016). For a Europe aiming to develop further its knowledge-based economy, these may be issues worth considering.

Therefore, almost 20 years later, not only the most important contribution management needs to make in the 21st century - to increase the productivity of knowledge work and knowledge workers (Peter F. Drucker, 1999) - is, apparently, yet to be accomplished, but also the challenge is increasing fast.

### **3.2 Core concepts**

Although Enterprise Search (ES) technologies date back to the late 1960s, when they were developed to search large online databases of scientific, commercial and legal information and to support the legal teams working on a number of large anti-trust suits in the USA - the breakup of AT&T being one example (White & Nikolov, 2013), the first true ES systems came to market approximately 20 years ago (Search Technologies, 2016).

In a more narrow perspective, the search performed by ES systems is confined to the enterprise repositories (White & Nikolov, 2013). However, in its classical sense, an ES system is a system providing search access over multiple and heterogeneous data sources (i.e. email attachments, dedicated content management servers, distributed file systems, private workspaces, web sites, intranets) to a variety of users whose needs are diverse



(Docurated, 2016; Martin Butler Research, 2009; Search Technologies, 2015), which is owned and controlled by an organization (Search Explained, 2015). In any case, ES systems are Information Retrieval (IR) systems, which have the primary goal to quickly provide users with the information they seek, particularly for important queries with high search frequency (Wu, Turpin, Thom, Scholer, & Wilkinson, 2014).

Effective ES systems allow users to convert distributed pieces of information into operational advantage (Martin Butler Research, 2009). ES technology relates users' queries to many different kinds of information in order to identify relevant, contextualized information and, in the process, perform light analysis (Gartner, 2015). Moreover, ES platforms can be used (White & Nikolov, 2013):

- For quick-changing data, and for ad hoc access to current data, often in real time or near real time;
- For better, more complete understanding of the organization and its business across information silos of both structured and unstructured information;
- For better customer understanding and service, by merging external social media information streams like Twitter or Facebook with internal customer records;
- In healthcare, to find a unified view of the patient;
- For faster decision support and strategic decision making;
- For more flexibility in delivering information depending on the question or context: customers and employees receive more relevant information because these systems understand who they are, what their role is, and what their question is.

A related concept to ES systems is that of Federated Search (FS). FS enables users to search across multiple repositories, each with its own search application, and be presented with an integrated and ranked list of results (White & Nikolov, 2013). It is common for modern ES systems to entail this feature.

Another concept related to ES systems is that of Unified Information Access platforms (UIA). These provide a single point of access to multiple heterogeneous sources of information. They are highly scalable and typically include tools for semantic understanding, including fuzzy matching and a range of search and text analytics outlines, as well as structured data and analytics operations. They are designed to work in the real-time or near-real-time updating and analytics, and also to combine elements of databases, business intelligence, and search technologies to make information access dynamic and ad hoc for business users (White & Nikolov, 2013). A number of search vendors are now planning to expand the scope of their products and services beyond searching through unstructured data through Unified Information Access Applications (UIA) (White & Nikolov, 2013).

Finally, a topical search engine is an engine that focuses on a particular topic. It covers a part of the whole Web rather than a particular website. Sometimes search terms can be ambiguous or have a different meaning depending on the context. By including only high quality, relevant sites in your engine, you narrow down the search domain and therefore make the results more precise and meaningful (Google, 2016).

Topical CSEs are a very valuable way of spreading the knowledge in a particular area and offer a tremendous value for users interested in the same topic. Through creating and grooming a well-curated index of sites, helping the user form the right query for a given use case and customizing the results, a topical engine can make finding the right information at the right time both pleasant and efficient (Google, 2016).

An example of topical CSEs could be Kritikos, a search engine developed by the Engineering Department at Liverpool University, where the search results are overlaid with additional data coming from the Learning Registry (Liverpool University, 2016).

Last, but not least, the global enterprise search market is expected to reach USD 8.90 billion by 2024, and Google, Inc. (Google Search Appliance), HP Autonomy (Verity),

SharePoint Search (Acquired by Microsoft), and IBM corporation are the leading players dominating it. However, Microsoft Corporation has recently stopped the commercialization of standalone products, namely 'Fast' and 'Transfer' (Grand View Research, 2016).

### 3.3 Typical features

The typical features that can be found in ES systems are the following:

- Allow users to search through multiple repositories from a single easy to use search interface (TATA Consultancy Services, 2015);
- Provide the ability to automatically tag documents or knowledge artefacts leveraging text mining and natural language processing algorithms, and eliminate dependency on manual processing (TATA Consultancy Services, 2015);
- Summarize results in the form of a concise snapshot (TATA Consultancy Services, 2015).
- Enable search not only through content management systems but also through other data sources, including enterprise data warehouses, email repositories, file systems, collaboration platforms, and more (TATA Consultancy Services, 2015);
- Allow search through internal information sources (includes transactional databases, email, instant messaging, documents, scanned documents, intranets and even some multimedia files such as voice and graphics) (Martin Butler Research, 2009);
- Allow search on other organizations information sources (the search in external sources is proliferating most rapidly). Organizations may wish to interrogate subsets of each other's information sources when collaborating with suppliers and customers, for example. Outside of this we have web sites and the new social media phenomena such as Facebook and Twitter. Some organizations are already interrogating Facebook as an additional resource in their recruitment process (Martin Butler Research, 2009);
- Consider user context such as role, location, and business group, while searching knowledge repositories (TATA Consultancy Services, 2015);
- Incorporate personal information into the results process to improve greatly the search results (Martin Butler Research, 2009);
- Multi-language search cross language information retrieval (CLIR) - describe the retrieval of information written in one language based on a query expressed in another (e.g. typing a query in English to retrieve documents written in Finnish) (White & Nikolov, 2013).

### 3.4 Expected benefits

In general, the benefits of ES are obvious, and boil down to basic issues of efficiency, productivity, cost avoidance, and the support of better-informed decision-making (Search Technologies, 2016). However, empirical studies which can support these claims should be further developed, since although IR systems, whether web-scale search or at an enterprise level, have a large impact on daily life, that impact is rarely measured. (Wu et al., 2014).

Some of ES systems expected benefits are presented in the literature as follows:

- Providing the right cross-repository enterprise search tool is fundamental to productivity and effectiveness (AIIM Industry Watch, 2014);
- To improve the operational efficiency is propelling the adoption of enterprise search solutions (Grand View Research, 2016);
- See their personal efficiency improve, through finding information more rapidly, all day, every day (Search Technologies, 2016);
- Time-saving data search capabilities (Grand View Research, 2016);
- Find information directly, rather than having to use their colleagues' time to broker questions, the answers to which are already well documented, somewhere (Search Technologies, 2016);

- Search for relevant information while having to go through each 'floor' (e.g. different applications) impedes the search process when time is of essence in decision making (TATA Consultancy Services, 2015);
- Day-to-day decisions are better informed, because of easier and more transparent access to supporting information (Search Technologies, 2016);
- An incremental improvement in the productivity of an organization's staff, across departments (Search Technologies, 2016);
- The ability to easily access, search, analyse, slice-and-dice, and universally exploit the growing mountain of information (Search Technologies, 2016);
- The value of enterprise search lies in its ability to provide simple, intuitive ways to consolidate information sources, whether structured or unstructured, and surface Information assets in the correct context and at the right time (Earley Information Science, 2016);
- Can considerably reduce the time and effort spent in searching and understanding the content (TATA Consultancy Services, 2015);
- Helps personnel better utilize their time to generate insights from search results rather than merely identifying relevant information (TATA Consultancy Services, 2015);
- Increased transparency, trust, and collaboration among knowledge teams (TATA Consultancy Services, 2015);
- Value at every stage of the business cycle (TATA Consultancy Services, 2015);
- Significant competitive advantage (TATA Consultancy Services, 2015);
- The sheer volume of information sources is causing serious inefficiencies to occur in the way we manage our time and resources (Martin Butler Research, 2009);
- Search provides improved visibility across diverse data sources and applications, acceleration of business processes, timely access to relevant data points, and the ability to analyse complex situations more effectively information search is the primary mechanism for creating value from information. As information sources proliferate so search will take on much greater importance (Martin Butler Research, 2009);
- Based strictly on the value of time saved, individuals in our study—that is, individual information seekers and content creators, consumers, and entrepreneurs—earn an ROI of 10:1 on average (Bughin et al., 2011);
- Enterprises earn still more, with an ROI of 17:1 as a result of time saved (Bughin et al., 2011);
- Despite the clear, measurable benefits of search to the economy, it would be a mistake to think about search only in terms that are easy to quantify. For example, search helps people find information in times of emergencies and helps them seek out people with similar interests—perhaps a support group for those coping with disease. Search also shifts the balance to empower individuals or small organizations with something to share that would otherwise reach only a small audience. None of these types of benefits may be easy to measure, but they are powerful nevertheless (Bughin et al., 2011);
- Most literature to date has looked at and quantified only three ways in which search creates value: by saving time, increasing price transparency, and raising awareness (Bughin et al., 2011);
- Benefits for government: better matching, time saved and raised awareness. Search helps customers, individuals, and organizations find information that is more relevant to their needs. Search accelerates the process of finding information, which in turn can streamline processes such as decision making and purchasing. Finally, search helps all manner of people and organizations raise awareness about themselves and their offerings, in addition to the value of raised awareness from an advertiser's perspective that has been the focus of most studies (Bughin et al., 2011);
- Search-enabled productivity gains enjoyed by knowledge workers in enterprise were worth up to \$117 billion in 2009 in the five countries studied. The figures ranged from \$49 billion to \$73 billion in the United States to \$3 billion to \$4 billion in Brazil (Bughin et al., 2011);

- Able to search more quickly and more easily than before will provide increasingly relevant results (Bughin et al., 2011);
- Subjects spent around one minute with the original lists compared to half a minute with the re-ranked lists on average (Wu et al., 2014);
- Of surveyed users saved 11% to 25% or more of time by using HPE IDOL solutions on search for information and gather business insights (Hewlett Packard Enterprise, 2016);
- AstraZeneca's app store has led to a better functional information discovery system and cross-departmental acceptance and use of better information discovery methods, which has improved its science, R&D productivity, and time to market. (Solinger & Schubmehl, 2016).
- As information sources proliferate, so the need to present a single search interface will become more important (Martin Butler Research, 2009)

### 3.5 Challenges

There are many challenges inherent to the adoption of ES solutions. One of them is the lack of awareness on the topic from decision makers (White & Nikolov, 2013). The core concepts around search and how leveraging those concepts in their specific circumstances will solve the problems are not well understood (Search Technologies, 2016). Moreover, the functionality of enterprise search applications and the benefits that effective search can have for the enterprise are also not well known, which constitutes a barrier to making a business case (White & Nikolov, 2013).

In part this is due to a couple of reasons. The first one is that the topic has not been researched in depth (White & Nikolov, 2013). Typically organisations usually have no research which identifies the most important tasks carried out by employees (White & Nikolov, 2013) and there is little academic research being carried out on the specific issues of enterprise search (White & Nikolov, 2013). Moreover, very few studies examine the benefit to the stakeholders whose interests are being served by the entire system (Wu et al., 2014). For example, the following questions raised (Seddon, P.B., Staples, S., Patnayakuni, R., & Bowtell, 1999) more than a decade ago remain unanswered. By being equipped with an IR system, are individuals better off and, if so, by how much? How much do organizations benefit by investing in an IR system? How much better off is a nation or society that uses IR tools? (Wu et al., 2014). So far, one study (Feldman, S., Duhl, J., Marobella, J.R., & Crawford, 2005) has been identified that aims to quantify the benefit of providing a better search engine. (Wu et al., 2014)

Furthermore, up until 2013 there has been no conference in the EU devoted to exploring the benefits and challenges of enterprise search, which could have formed a community of interest around the topic (White & Nikolov, 2013). Fortunately, from that moment on, Enterprise Search Europe has been the leading conference focusing on Enterprise Search in Europe (Search Explained, 2015).

The second reason is the difficulty in measuring the costs and benefits, hence in building a business case (White & Nikolov, 2013). There is no sound approach to measure the costs and benefits associated with using these information resources (Martin Butler Research, 2009; Search Technologies, 2016) and this is a key issue (Search Technologies, 2016). Additionally, the difficulty in measuring a general productivity improvement, remains a problem for many decision makers (Search Technologies, 2016).

Consequently, it is very hard to justify such initiatives. The good news is that investments in targeted search-based applications, deployed in environments where the benefits are easier to measure, are more likely to be signed-off (Search Technologies, 2016). This requires a "search story" prepared, complete with evidence and justification for its place within business applications in the organization (Search Technologies, 2016). If there is a valid and credible story to tell, then it is necessary to get the details of the business problem and lay out a concise, realistic plan of attack (Search Technologies, 2016).

Additionally, given that most organisations already perform search in various ways, mostly through other enterprise applications and only for specific processes, the value

may not be fully appreciated (White & Nikolov, 2013). Usually the view is that google is the best search application and offers what is needed (White & Nikolov, 2013) and also that enterprise portals (e.g. sharepoint) offer the required enterprise search functionality (White & Nikolov, 2013). Also, often information is not seen as a business asset (White & Nikolov, 2013), which makes ES to be regarded as a low-priority investment (White & Nikolov, 2013).

One of the major challenges for ES solutions is to provide an integrated and value-adding retrieval of both structured and unstructured information and merge them into a sort of 'Hybrid structured data'. The vision is thus of Unified Information Access, which would provide the end-user with a user-friendly interface capable of retrieving heterogeneous data sources and providing added value by making use of semantic modelling (White & Nikolov, 2013). But this remains a tough subject to get right (Search Technologies, 2016) especially because of the substantial technological complexity it entails.

Some of this complexity results from its federated search capability, which requires: a) Considerable flexibility in combining results in different ways; b) Maintaining and mapping security credentials across applications which may not have a common identity management application; c) Advanced detection and removal of duplicate documents, which has long been a major challenge for search vendors; d) Combining results-list and page navigators and e) Translating queries into different search syntaxes (White & Nikolov, 2013). Additionally, although most enterprise search vendors offer some degree of federated search, the performance of federated search applications is very dependent on content quality and metadata quality (White & Nikolov, 2013). Therefore, the key to successful ES for many organizations will be simplicity (Martin Butler Research, 2009).

The challenge posed by the complexity of ES enters into a new level when faced with the shortage of skills to do it. There is a lack of internal expertise to support the implementation (White & Nikolov, 2013) and there is a shortage of skilled professionals to join search vendors as development and implementation engineers, and to join enterprise search support teams (White & Nikolov, 2013). This also leads to the lack of support post-implementation, or the lack of a search support team which can cope with the changing business requirements. (White & Nikolov, 2013).

Managing the users expectations is also a challenge, since they may be disappointed by the results a search delivers (Martin Butler Research, 2009). Actually, a common flaw in the use of information search technologies is overconfidence in the results (Martin Butler Research, 2009). While getting search technologies to do the donkey work is fine, expecting these same technologies to determine meaning and relevance is folly, and will result greater costs in the long run (as we make erroneous decisions and supply various authorities with flawed information) (Martin Butler Research, 2009).

Overall, the implementation of ES technologies is still a risky initiative, that it is of potential (White & Nikolov, 2013) value to most, if not all, employees, but no single department wishes to take responsibility for making a business case (White & Nikolov, 2013).

## **4. Proposed approach**

Considering the previous discussion on ES technologies, we conclude that they are adequate to address some of the challenges presently faced by the KCMD and fulfil the KB requirements outlined earlier. Consequently, this feasibility study is about the implementation of an ES system for the KCMD, which we have named Knowledge Browser (KB) and, therefore, all general ES benefits and challenges previously discussed should also be considered both in this study and in its following developments.

Like any other system, the KB may be implemented in different ways, any of which involves resources, takes time and entails risks. The purpose of this section is to outline

some of the possible approaches for developing it and to identify the best one. In this regard, we will start by describing the different alternatives considered, and then explain which is the preferred one, and why.

From the perspective of an organisation, there are seven ways in which an enterprise search application can be procured and implemented, as follows (White & Nikolov, 2013).

**For commercial products:**

1. An incumbent systems integrator which is already providing support for enterprise applications (e.g. IBM or Accenture) could also provide an enterprise search application;
2. An integration company that specialises in search applications could be asked to select and implement an enterprise search application;
3. The organisation could work directly with the enterprise search vendor, who would then provide not only the software but also professional services support for the implementation;
4. The search application may be embedded in another application on an OEM basis, where often the company developing the search software is not identified;

**In the case of open source products:**

5. The open source application is purchased as a product from a specialist integrator, where much of the development work has already been carried out and in some cases proprietary code has been integrated with the open-source code (e.g. Attivio, IntraFind);
6. The organisation could decide to develop the application using only internal staff resources;
7. A specialist developer could be used to provide a fully customised application, though this is often carried out in conjunction with internal resources. Within the organisation there will be a requirement for both IT support to ensure that the hardware and software applications are working to agreed technical performance standards, but there will be a much greater requirement for a search support team.

Presently, the decision is to make an open source implementation with resources internal to the KCMD (option 6). Still, this scenario implies considering additional options, which are a) development from scratch, b) reuse of functionally relevant OSS components and c) reuse of functionally relevant systems.

## **4.1 Development from scratch**

This approach is the one where all core software is developed from scratch. Building on common COTS<sup>5</sup> or OSS<sup>6</sup> components (e.g. databases, web servers), the functional requirements are designed and coded from zero, without reusing any component that could partially or completely fulfil their functionality.

This approach gives us (almost) total freedom in regard to the technologies used (e.g. programming languages), hence we can reuse the skills of the existing development team. On the other hand, it is likely the approach that will take more time to develop, because of its extremely low rate of software reuse.

Upon development, maintaining it will also become a very specialized task, creating a high dependency from the developers involved. One major risk is that other alternatives might exist which, in spite of the efforts made during development, perform better, hence affecting the relevance of the solution and its cost-effectiveness.

---

<sup>5</sup> Commercial off-the-shelf

<sup>6</sup> Open source software

## 4.2 Reuse of functionally relevant OSS components

This approach is the one where one or more OSS components are reused to, partially or fully, implement the requirements. Before designing the solution it is necessary to understand which components qualify in this regard, choose the best ones, and design the rest of the solution around them.

This approach implies that we use (to a certain extent) some of the technologies inherent to the OSS components adopted (e.g. programming languages, databases), which might imply an adjustment of the development team to the skills required by those products. The fact that tested and proven software is being reused will decrease the time to develop the solution and reduce the inherent maintenance effort, while creating less dependencies from the developers involved than in the previous alternative.

One of the most commonly used OSS components for implementing search engines such as the one herein envisioned is Open Semantic Search<sup>7</sup>. Open Semantic Search is free software for implementing a Search Engine, Explorer for Discovery of large document collections, Media Monitoring, Text Analytics, Document Analysis & Text Mining platform based on Apache Solr or Elasticsearch open-source enterprise-search and Open Standards for Linked Data & Semantic Web.

Elasticsearch<sup>8</sup> is an emerging and fastly becoming adopted technology, which is a distributed, RESTful search and analytics engine. Solr<sup>9</sup>, on the other hand, is highly reliable, scalable and fault tolerant, providing distributed indexing, replication and load-balanced querying, automated failover and recovery, centralized configuration and more. Solr powers the search and navigation features of many of the world's largest internet sites.

The tool which supports the KCMD data catalogue is also an OSS component named CKAN<sup>10</sup>. CKAN is a powerful data management system that makes data accessible by providing tools to streamline publishing, sharing, finding and using data. CKAN is also built on Solr; therefore, some of the present knowledge of the team which results from using CKAN can eventually be useful while developing solutions equally based on Solr.

## 4.3 Reuse of functionally relevant systems

Sometimes, complete systems do exist in an organization, which could be reused, with some adjustments, to meet the requirements of a different solution. This approach considers this option. Here the design should focus on which parts of the system have to be changed to support the different requirements.

This approach opens the possibility of reusing the existing knowledge and much of the existing software to develop the new system. Considering this would be a new modified instance of an existing system, this approach would be the one to go for achieving the fastest results. Since most of the software would be reused, this would also be the solution less prone to errors, hence the most reliable one (at least in the first months of operation). Moreover, in this scenario, the team involved would hardly require new skills.

In the JRC, only one system presently being used has been identified as potentially capable of being reused for meeting the requirements expressed. This is the European Media Monitor News Explorer<sup>11</sup>, which has already been briefly described in previous sections. Interestingly, there is even a case where the EMM has been significantly reused to deliver

---

<sup>7</sup> <https://www.opensemanticsearch.org/>

<sup>8</sup> <https://www.elastic.co/products/elasticsearch>

<sup>9</sup> <http://lucene.apache.org/solr/>

<sup>10</sup> <http://ckan.org/>

<sup>11</sup> <http://emm.newsexplorer.eu>



a new system with slightly different requirements. This is the case of the MEDISYS<sup>12</sup> system.

However, after a brief analysis of this possibility, we concluded that there are some limitations in this approach. Technically, the most relevant one is that the present solution is only able to handle small files (up to 100 KB). Additionally, some other features would have to be developed to meet the specific requirements herein discussed. Non-technical limitations, on the other hand, are determinant and have to do with the lack of personal to develop and maintain, in a sustainable way, such initiative. This limitation unfortunately impedes any developments, for the time being.

#### **4.4 Crosscutting issues**

There are also aspects to consider irrespectively of the approach followed. Firstly, a dedicated technical infrastructure (hardware and software) will have to be developed. Such an infrastructure will imply, in the best case, virtual servers which will consume the resources from the existing physical ones. In the worst case, however, new physical servers to support the new requirements might have to be acquired.

Secondly, in the best case, some training might be necessary to the people involved in the development and maintenance of the solution. In the worst case, contracting new people might be necessary, if all of the effort of the existing resources is already being used in other initiatives. Both the acquisition of material and the contracting of people usually require significant time.

Finally, the success of the implementation will also depend on the availability of the data sources of interest to be queried from the ES solution implemented. While most of external public data sources should not require specific technology; other, namely the Commission internal systems, will certainly require it. While, for example, Connected already provides interfaces which seem appropriate for this (Aeolian, 2016), the same kind of features in other relevant systems would have to be verified and documented.

#### **4.5 Proposal**

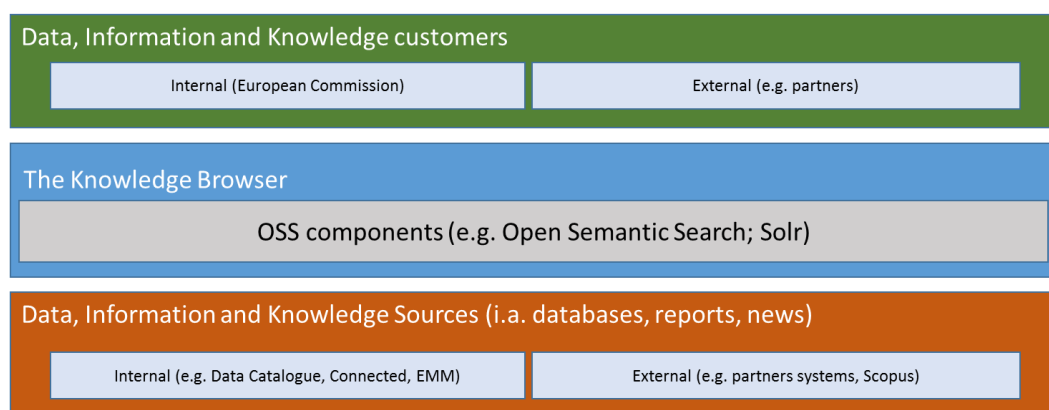
Considering all the above, we have chosen to adopt an approach based on functionally relevant OSS components, over the other two. As earlier said, unfortunately it is not possible to take the approach based on EMM, given essentially the limitations regarding the human resources available.

Therefore, when compared to the alternative of developing the whole system from scratch, the selected approach is the one which is able to provide results faster and less prone to failure, since it is based on well proven and widely used software. Moreover, by using such components, the dependency on specific developers is minimized and, if some of the technology used is already known by the existing team, then their knowledge in this regard will make the development and maintenance easier.

---

<sup>12</sup> <http://medisys.newsbrief.eu/medisys/homeedition/en/home.html>





**Figure 3: Architecture of the proposal**

The architecture of the proposed solution is depicted in Figure 3. The idea is that both internal (Commission including the KCMD) and external users can access the KB for retrieving the data, information and knowledge relevant for migration and demography policy making and support. As earlier said, the KB will be developed based on OSS components such as Open semantic Search and Solr, and will gather the relevant content among internal (e.g. data and information catalogues, Connected and EMM) and external (e.g. selected websites and Scopus).

Given the approach taken, most of the KB requirements could be already fulfilled out of the box, just by selecting an appropriate set of components. This means that most of the effort will be dedicated to the installation and configuration of such components, and only a small part of the effort is supposed to be dedicated to customized software development, for delivering requirements uncovered by the OSS components chosen.

With this in mind, and based on experience, developing an initial prototype of the KB which would implement most of the relevant requirements could take up to one semester. This would include designing the solution and implementing and testing it by following an iterative, interactive and incremental approach.

## 5. Evaluation

Now that the best approach for developing the KB has been selected, it becomes necessary to understand its costs and benefits. However, in many cases (if not all), this can be a complex task, because not all costs and benefits are tangible and clear, and the approaches to do it are diverse. Usually, a common strategy to overcome these challenges is to keep the calculations simple, in the sense that they become easy to verify and with the subjectivity reduced to a minimum. We will start by an overview of the approaches commonly used, and then proceed with the quantitative and qualitative evaluations of the proposed solution.

### 5.1 Evaluation of information retrieval systems

Typically, the evaluation of an IR (information retrieval) system is focused on its search engine component, measuring how quickly it can respond to a query, or how good the retrieved information is, relative to some set of relevance judgments. This is called a system-oriented evaluation and is centred on the retrieval effectiveness and efficiency of the search engine component of an IR system.

In this context, efficiency is typically measured in time and memory usage, both the resources required by the indexing and reprocessing subsystem of a search engine and the resources required to resolve the query. Effectiveness, on the other hand, is measured using the Cranfield methodology (Cleverdon, 1984), in which a set of queries, and a corpus of documents with known relevance scores relative to those queries, are used to compute

a score for a ranked list returned for each query. Typical measures include variations on precision and recall (Wu et al., 2014).

However, another way to evaluate an IR system is the user-oriented evaluation, which includes information seekers in the evaluation and measures factors such as how well a search system can help users achieve their goals, how well users perceive they are achieving their goals, or just straight out user satisfaction with the system (Belkin & Vickery, 1985).

User-oriented evaluation is usually conducted in a laboratory setting in which recruited subjects, who supposedly represent a user population, are asked to participate in a scenario that simulates a real-world search problem. Subpopulations of the users are given different systems to allow system comparison. Outcome measures include time to finish tasks, number of (relevant) documents found, number of aspects of a topic found, the user's perceived performance on the task, user satisfaction with a task, and measures of effort such as number of queries issued or number of mouse clicks (Over, 2001; Wilkinson & Wu, 2004).

The following three criteria are often used in a user-oriented evaluation (Wu et al., 2014): utility, satisfaction, and use. The utility criterion makes the assumption that an IR system ought to be evaluated based on how useful it is to an information seeker.

Satisfaction is a subjective criterion based on information seekers' reactions to a system. Similar to the utility approach, the satisfaction approach also moves away from evaluation of system performance only. It goes beyond an information seeker's perception of a document; it addresses an information seeker's perception of the whole information-seeking environment. Thus, satisfaction is a multifaceted measurement that may include a system's retrieval effectiveness, efficiency, and usability, how well it fulfils a search task, and so on (Wu et al., 2014).

Use is a criterion that considers the use to which the user puts the Information gained from the interaction. Unlike the above measures that are focused on how much an information unit or an information service is of perceived value to an information seeker, the use criterion considers how much the information gained from a search helps an information seeker to solve tasks and problems that require information for resolution. After all, information seeking is typically a means but not an end (Wu et al., 2014).

Both system-oriented and user-oriented evaluations of an IR system are essential for the development and advancement of such systems. However, another factor that should be included in the evaluation of an IR system is how the system impacts on the community or organization that the information seekers belong to, because this is an important dimension of information system success (DeLone & McLearn, 1992).

This issue becomes more important when an IR system is used in the context of an enterprise search, because the enterprise needs to understand its costs and benefits and quantify how much productivity can be improved if employees and customers are provided with an improved information service (Wu et al., 2014).

## **5.2 Quantitative evaluation of the proposed solution**

Sassone (Sassone, 1988) surveyed eight generic methodologies that quantify the cost and benefit of information systems: decision analysis, structural models, breakeven analysis, subjective analysis, cost displacement or avoidance, cost effectiveness analysis, time savings times salary, and the work-value model (Wu et al., 2014).

The time savings times salary (TSTS) is an increasingly popular methodology for estimating the value of office information systems. It does so by estimating the percentage of workers' time the system will save, and multiplying it by the workers loaded salaries or wages (Sassone, 1988). The two premises underlying the TSTS method are that:

- a) a worker's value to an enterprise equals their cost to the enterprise;
- b) saving X% of a worker's time is worth X% of the worker's cost.

Therefore, if an enterprise has  $Y$  workers each of which costs  $\$S$  to the organization, and a new information system is expected to save an average  $X\%$  of each worker's time, then the annual value of the system is  $X * Y * \$S$  (Wu et al., 2014).

To perform the quantitative part of the cost-benefit analysis of the proposed solution we have considered two scenarios. The first scenario represents the situation where the system is not developed as proposed, hence things continue as they are presently, without change. The second scenario, on the other hand, represents the situation where the proposed system is developed and adopted by the KCMD, hence changing the present processes.

For each of these scenarios we have then estimated the investment and operational costs based on human resources and equipment, and calculated the benefits based on the TSTS. Previously, we have seen that the time taken in searching activities can vary from 12% to 22%. Moreover, based on the pilot project on climate change and migration, we estimate that the time taken to build such a catalogue can take between 5% and 15% of each person involved. We have also seen that the time saved by introducing ES solutions can also vary significantly from 11% to 50%. Therefore, for each of the scenarios we have developed an optimistic (21% time taken in general searches; 12% time taken in the searches for the catalogue; 40% time saved with ES) and a conservative approach (13% time taken in general searches; 9% time taken in searches for the catalogue; 20% time saved with ES), as follows.

There are also some elements relevant for the analysis which are independent of the scenario taken. These are presented in Table 1.

**Table 1: Structural elements**

	Quantity	Unit
KMD experts involved in the processes	20	pm
IT experts involved in developing the KB	1	pm
IT experts involved in maintaining the KB	0,5	pm
Monthly cost of an expert (generic)	6000	euro
Server cost (generic)	10000	euro
Estimated duration of the KB development	6	months

### **5.2.1 Scenario 1: the proposed solution is not developed**

This scenario represents the present situation and assumes that the ES solution will not be adopted. As earlier said, for each scenario we will consider a conservative and an optimistic approach.

Scenario 1A, represents the conservative approach. In Table 2 we can see the detailed effort and benefits of this approach, where we consider 13% as the time taken per person in the "retrieve" process and 9% in the "gather and share" process. Since no ES solution will be developed, there are no benefits; hence, the overall time spent in both processes by each person is 22%.

**Table 2: Effort and benefits breakdown. No development. Conservative approach.**

<b>Scenario 1A (no development - conservative)</b>	<b>Quantity</b>	<b>Unit</b>
Time taken in search activities before introducing the ES solution	13 %	
Time saved by introducing ES solution	0 %	
Time taken in search activities after introducing the ES solution	13 %	
Time taken in catalogue activities before introducing the ES solution	9 %	
Time saved by introducing ES solution	0 %	
Time taken in catalogue activities after introducing the ES solution	9 %	
<b>Total time taken</b>	<b>22 %</b>	

In Table 3 we can see the cost breakdown of this approach, considering a 5 year timespan. Since no developments will be made, there are no investment costs, nor maintenance costs associated. According to this conservative approach, every year the KCMD will spend, searching information for everyday activities or to specifically build the catalogue the equivalent to approximately 2.64 months of each persons' time, or 4.4 persons.year, if we consider all KMD experts involved.

**Table 3: Cost breakdown. No development. Conservative approach.**

	<b>CAPEX</b>	<b>OPEX</b>					
	<i>Year 1</i>	<i>Year 1</i>	<i>Year 2</i>	<i>Year 3</i>	<i>Year 4</i>	<i>Year 5</i>	<i>Total</i>
KMD experts	0,00 €	316.800,00 €	316.800,00 €	316.800,00 €	316.800,00 €	316.800,00 €	1.584.000,00 €
IT experts	0,00 €	0,00 €	0,00 €	0,00 €	0,00 €	0,00 €	0,00 €
Hardware	0,00 €	0,00 €	0,00 €	0,00 €	0,00 €	0,00 €	0,00 €
<b>Total</b>	<b>0,00 €</b>	<b>316.800,00 €</b>	<b>316.800,00 €</b>	<b>316.800,00 €</b>	<b>316.800,00 €</b>	<b>316.800,00 €</b>	<b>1.584.000,00 €</b>

Scenario 1B, represents the optimistic approach. In Table 4 we can see the detailed effort and benefits of this approach, where we consider 21% as the time taken per person in the "retrieve" process and 12% in the "gather and share" process. Since no ES solution will be developed, there are no benefits; hence, the overall time spent in both processes by each person is 33%.

**Table 4: Effort and benefits breakdown. No development. Optimistic approach.**

<b>Scenario 1B (no development - optimistic)</b>	<b>Quantity</b>	<b>Unit</b>
Time taken in search activities before introducing the ES solution	21 %	
Time saved by introducing ES solution	0 %	
Time taken in search activities after introducing the ES solution	21 %	
Time taken in catalogue activities before introducing the ES solution	12 %	
Time saved by introducing ES solution	0 %	
Time taken in catalogue activities after introducing the ES solution	12 %	
<b>Total time taken</b>	<b>33 %</b>	

In Table 5, we can see the cost breakdown of this approach, considering a 5 year timespan. Since no developments will be made, there are no investment costs, nor maintenance costs associated. According to this optimistic approach, every year the KCMD will spend, searching information for everyday activities or to specifically build the catalogue the equivalent to approximately 3.96 months of each persons' time, or over 6.6 persons.year, if we consider all KMD experts involved.

**Table 5: Cost breakdown. No development. Optimistic approach.**

	CAPEX	OPEX					
	Year 1	Year 1	Year 2	Year 3	Year 4	Year 5	Total
KMD experts	0,00 €	475.200,00 €	475.200,00 €	475.200,00 €	475.200,00 €	475.200,00 €	2.376.000,00 €
IT experts	0,00 €	0,00 €	0,00 €	0,00 €	0,00 €	0,00 €	0,00 €
Hardware	0,00 €	0,00 €	0,00 €	0,00 €	0,00 €	0,00 €	0,00 €
<b>Total</b>	<b>0,00 €</b>	<b>475.200,00 €</b>	<b>475.200,00 €</b>	<b>475.200,00 €</b>	<b>475.200,00 €</b>	<b>475.200,00 €</b>	<b>2.376.000,00 €</b>

## 5.2.2 Scenario 2: the proposed solution is developed and adopted

This scenario represents the situation where the ES solution is developed. Likewise, for each scenario we will also consider a conservative and an optimistic approach.

Scenario 2A, represents the conservative approach. In Table 6 we can see the detailed effort and benefits of this approach, where we consider 13% as the time taken per person in the "retrieve" process and 9% in the "gather and share" process. Since the ES solution will be developed, and this approach is conservative, we have considered the time saved by introducing the solution to be 20%; hence, the overall time spent in both processes by each person may become 17.6%.

**Table 6: Effort and benefits breakdown. Development. Conservative approach.**

<b>Scenario 2A (development - conservative)</b>	<b>Quantity</b>	<b>Unit</b>
Time taken in search activities before introducing the ES solution	13 %	
Time saved by introducing ES solution	20 %	
Time taken in search activities after introducing the ES solution	10,4 %	
Time taken in catalogue activities before introducing the ES solution	9 %	
Time saved by introducing ES solution	20 %	
Time taken in catalogue activities after introducing the ES solution	7,2 %	
<b>Total time taken</b>	<b>17,6 %</b>	

In Table 7 we can see the cost breakdown of this approach, considering a 5 year timespan. Since the ES solution will be developed, there are costs in the first year to consider (CAPEX), related to the IT expertise and hardware required. Overall, the solution's TCO<sup>13</sup> is of 226.000,00 euros.

Additionally, we have also to consider the inherent yearly maintenance costs associated to the IT expertise required. According to this conservative approach, every year the KCMD will spend, searching information for everyday activities or to specifically build the catalogue, supported by the ES system, the equivalent to approximately 2.1 months of each persons' time, or 4.0 persons.year, if we consider all persons involved.

**Table 7: Cost breakdown. Development. Conservative approach.**

	CAPEX	OPEX					
	Year 1	Year 1	Year 2	Year 3	Year 4	Year 5	Total
KMD experts	0,00 €	253.440,00 €	253.440,00 €	253.440,00 €	253.440,00 €	253.440,00 €	1.267.200,00 €
IT experts	36.000,00 €	36.000,00 €	36.000,00 €	36.000,00 €	36.000,00 €	36.000,00 €	216.000,00 €
Hardware	10.000,00 €	0,00 €	0,00 €	0,00 €	0,00 €	0,00 €	10.000,00 €
<b>Total</b>	<b>46.000,00 €</b>	<b>289.440,00 €</b>	<b>289.440,00 €</b>	<b>289.440,00 €</b>	<b>289.440,00 €</b>	<b>289.440,00 €</b>	<b>1.493.200,00 €</b>

Scenario 2B, represents the optimistic approach. In Table 8 we can see the detailed effort and benefits of this approach, where we consider 21% as the time taken per person in the "retrieve" process and 12% in the "gather and share" process. Since the ES solution will be developed, and this solution is optimistic, we have considered the time saved by

<sup>13</sup> TCO = investment costs – operation costs

introducing the solution to be 40%; hence, the overall time spent in both processes by each person may become 19.8%.

**Table 8: Effort and benefits breakdown. Development. Optimistic approach.**

<b>Scenario 2B (development - optimistic)</b>	<b>Quantity</b>	<b>Unit</b>
Time taken in search activities before introducing the ES solution	21 %	
Time saved by introducing ES solution	40 %	
Time taken in search activities after introducing the ES solution	12,6 %	
Time taken in catalogue activities before introducing the ES solution	12 %	
Time saved by introducing ES solution	40 %	
Time taken in catalogue activities after introducing the ES solution	7,2 %	
<b>Total time taken</b>	<b>19,8 %</b>	

In Table 9, we can see the cost breakdown of this approach, considering a 5 year timespan. Since the ES solution will be developed we have to consider the same TCO as before.

According to this approach, every year the KCMD will spend, searching information for everyday activities and to specifically build the catalogue, the equivalent to approximately 2.4 months of each persons' time, or over 4.46 persons.year, if we consider all KMD experts involved.

**Table 9: Cost breakdown. Development. Optimistic approach.**

	CAPEX		OPEX				
	Year 1	Year 1	Year 2	Year 3	Year 4	Year 5	Total
KMD experts	0,00 €	285.120,00 €	285.120,00 €	285.120,00 €	285.120,00 €	285.120,00 €	1.425.600,00 €
IT experts	36.000,00 €	36.000,00 €	36.000,00 €	36.000,00 €	36.000,00 €	36.000,00 €	216.000,00 €
Hardware	10.000,00 €	0,00 €	0,00 €	0,00 €	0,00 €	0,00 €	10.000,00 €
<b>Total</b>	<b>46.000,00 €</b>	<b>321.120,00 €</b>	<b>321.120,00 €</b>	<b>321.120,00 €</b>	<b>321.120,00 €</b>	<b>321.120,00 €</b>	<b>1.651.600,00 €</b>

### 5.2.3 Comparative analysis

To understand if the ES solution is worthwhile in the case of the KCMD, we will make two different comparisons. First, we will compare the two conservative approaches, in the circumstances where the ES solution is and is not developed. Then, we will compare the two optimistic approaches, for the same circumstances.

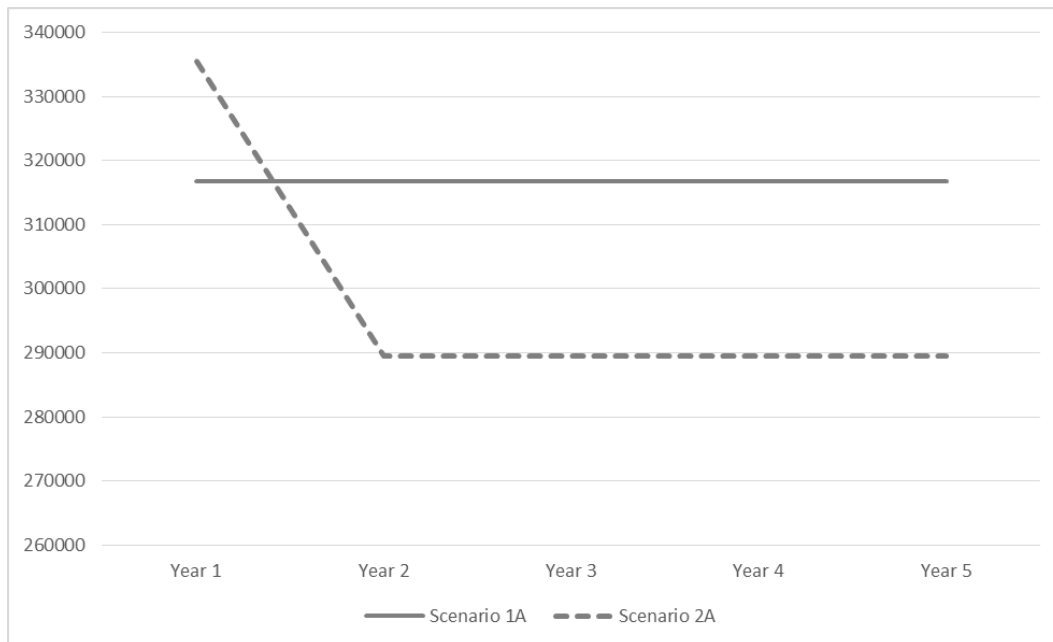
#### Conservative perspective

From the conservative perspective, we can see, from the cost evolution of both scenarios depicted in Figure 4, that there is quantitative benefit of implementing the ES solution. While in the first year the costs of implementing the solution exceed those of not implementing it, as expected, in the following years the costs of using the ES solution are smaller than those of not using it.

Specifically, after the first year, the benefits of using the solution are those equivalent to approximately 0.54 months of each persons' time and 0.4 person.year, if all KMD experts involved in the processes are considered. After five years, the financial benefit is around 18.160 euros, which is not significant.

Overall, the ROI<sup>14</sup> of this perspective is roughly of 1:1, well out of the range [10:1; 17:1] verified in similar initiatives (Bughin et al., 2011). Therefore, we find this perspective very unlikely to occur.

<sup>14</sup> ROI = ( benefit – investment ) / investment

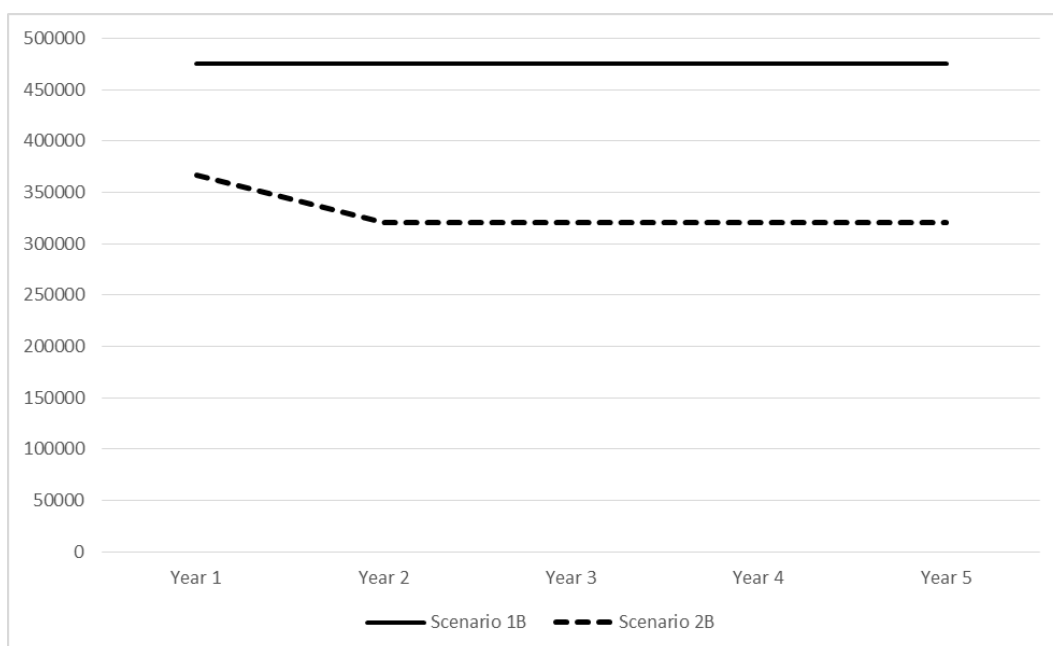


**Figure 4: Costs. Conservative perspective scenarios.**

### Optimistic perspective

In the optimistic perspective, we can see, from the cost evolution of both scenarios depicted in Figure 5, that there is a significant benefit of implementing the ES solution. This is true even in the first year, during which the costs of using the ES solution are increased by the fact that it is being implemented. Then, in the following years the costs of not using the ES solution are bigger than those of using it.

Specifically, after the first year, the benefits of using the solution are those equivalent to approximately 1.56 months of each persons' time and 2.14 person.year, if all KMD experts involved in the processes are considered. After five years, the financial benefit is around 724.400 euros, which is already significant.



**Figure 5: Costs. Optimistic perspective scenarios.**



Overall, the ROI of this perspective is roughly of 15:1, on the superior limit of the range [10:1; 17:1] verified in similar initiatives (Bughin et al., 2011). Therefore, we find this perspective more likely to occur than the conservative one.

### 5.3 Qualitative evaluation of the proposed solution

But not all costs and benefits that result from introducing an information system into an organization are easily quantifiable, which means they cannot usually be assessed in a quantitative way. However, these outcomes must, in any case, be considered in the overall assessment, so that we can have the most complete picture to support decision making.

In our particular proposal we do not see relevant intangible costs. However, we do see the following relevant intangible benefits which we assess, in a qualitative way, as follows:

- **Relevance of the KCMD and of the JRC:** the fact that users external to the KCMD and the JRC might use the KB for their own activities has the potential to increase the KCMD and JRC's visibility and relevance, inside and outside the Commission;
- **Third-party performance:** the fact that users external to the KCMD and the JRC (either from the Commission or not) may use the KB to develop their own activities, may also bring benefits to their own processes, similarly to what is expected to happen with the KCMD processes; therefore, the KB may contribute to enhance the performance of 3<sup>rd</sup> parties;

**Motivation of researchers and other knowledge workers:** the persons involved in processes such as the ones previously described are typically highly skilled workers, and some of the tasks the KB is supposed to support are not so much intellectually demanding. For such workers, performing such tasks for a long period of time can be very boring and contribute to their demotivation, which will then negatively impact the rest of the tasks they're involved in. By being supported by the KB, the motivation of these workers is, at least, expected to increase for not spending so much time doing unwanted tasks.

## 6. Conclusions and recommendations

Improving the way in which data, information and knowledge are gathered and shared is a priority for the Commission, the JRC and the KCMD. Moreover, considering the huge effort such activities will require, as already experienced in a pilot project, the scarce resources available at the KCMD for this purpose and the fact that, in foreseeable future this situation is unlikely to change, it is urgent that technology is developed to support overcoming these challenges.

With this in mind, we have analysed the present processes of the KCMD involved in those activities. We have identified their challenges, which were then used to determine opportunities for improvement. Then, we proposed the development of an Enterprise Search solution – the Knowledge Browser (KB) – to address some of those opportunities and defined its general requirements.

Afterwards, we have outlined three different approaches for developing it, as well as some crosscutting aspects to consider, regardless of the approach taken. The first approach was to develop the KB from scratch, the second to develop it based on OSS components, and the third to reuse an existing JRC system – the EMM. The approach selected was the second, considering especially the limitations in human resources that impede us from reusing the EMM, as well as the effort required to develop a solution from scratch, and the risk of developing something minor, when compared to existing solutions, hence hampering the cost-effectiveness of the overall solution.

Then, we briefly described the proposed approach, pointing out already some OSS components that could be used for the purpose, and estimated that one semester could



be enough to design and develop the KB as proposed, implementing most of its relevant requirements and providing the necessary resources (human and material) are available in due time.

Finally, we have evaluated the proposed solution qualitatively and quantitatively. The qualitative benefits are, in a nutshell, the enhanced relevance of the KCMD and the JRC, the enhanced performance of partners (inside and outside the Commission), and the increased motivation of the KCMD and the JRC researchers and other knowledge workers.

To ascertain the quantitative benefits of the KB we have adopted two perspectives. One conservative and another optimistic. While from a conservative perspective, these benefits are not so significant; from an optimistic perspective, the quantitative benefits are already significant, allowing the KCMD to save approximately 1.56 months of work per year of each person or 2.14 person.year if all KMD experts involved are considered. These savings are of over 150.000 euros per year and, consequently, over 750.000 euros after five years.

Based on the ROI of related studies, the most likely outcome will be closer to the optimistic than to the conservative perspective. The case study used (the KCMD) is very limited in scope. Therefore, much bigger benefits can be expected if we widen the scope of the utilization of the KB. For example, the JRC alone has 50 knowledge units (12 knowledge management and 38 knowledge production), which could benefit from the KB up to 7.5 million euros per year. These benefits would be even bigger for the Commission if the KB would be made available to policy officers.

Consequently, we recommend to proceed with the design and development of a prototype of the KB, based on the proposed approach, since we consider that the benefits (qualitative and quantitative) it may bring to the KCMD (and potentially to the JRC and the Commission) will have short and long term impacts in its efficiency, effectiveness and relevance that will clearly outperform its costs.

In this study, we have analysed the feasibility of developing the KB with internal resources and based on open source solutions. Nonetheless, it would be interesting to understand the cost-benefit of developing it with alternative commercial solutions as well.

It has been reported and confirmed by our analysis of the state of the art on ES solutions that there are not so many empirical studies which state, in quantitative terms, the benefits of this technology. Therefore, an ex-post evaluation of the system implementation at the KCMD (or at the JRC) could have scientific value as well.

Finally, assuming the challenges addressed by the KB exist in many other units of the JRC, it is recommended that, once the prototype is completed, and assuming the solution costs and benefits are more clear, the initiative is sufficiently divulgated, in order to avoid duplicate efforts, gather synergies and collect contributions which may increase its likelihood of becoming a reference solution and being widely adopted by the JRC and used further in the Commission and by other external stakeholders.



## 7. References

- Aeolian, M. (2016). Some JIVE REST API examples for extracting data from Connected. Retrieved December 7, 2016, from <https://connected.cnect.cec.eu.int/people/aeolima/blog/2016/04/05/some-jive-rest-api-examples-for-extracting-data-from-connected>
- AIIM Industry Watch. (2014). Search and Discovery – Exploiting Knowledge, Minimizing Risk. Retrieved November 18, 2016, from [http://www.aiim.org/Resources/Research/Industry-Watches/2014/2014\\_Sept\\_Search-and-Discovery](http://www.aiim.org/Resources/Research/Industry-Watches/2014/2014_Sept_Search-and-Discovery)
- Bughin, J., Corb, L., Manyika, J., Nottebohm, O., Chui, M., Barbat, B. de M., & Said, R. (2011). *The impact of Internet technologies: Search*. Retrieved from [https://www.google.it/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEWjl3OzGgZzQAhUJWj4KHts\\_Bb8QFggiMAA&url=https%3A%2F%2Fwww.mckinsey.com%2F~%2Fmedia%2Fmckinsey%2Fdotcom%2Fclient\\_service%2FHigh%2520Tech%2FPDFs%2FImpact\\_of\\_Internet\\_tech](https://www.google.it/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEWjl3OzGgZzQAhUJWj4KHts_Bb8QFggiMAA&url=https%3A%2F%2Fwww.mckinsey.com%2F~%2Fmedia%2Fmckinsey%2Fdotcom%2Fclient_service%2FHigh%2520Tech%2FPDFs%2FImpact_of_Internet_tech)
- Cleverdon, C. (1984). Optimizing convenient online access to bibliographic databases. *Information Services and Use*, 4(1-2), 37-47.
- Crabtree, R. A., Fox, M. S., & Baid, N. K. (1997). Case Studies of Coordination Activities and Problems in Collaborative Design. *Research in Engineering Design*, 9, 70-84. Retrieved from <http://www.eil.utoronto.ca/wp-content/uploads/kbd/papers/crabtree-red97.pdf>
- Davenport, T. H. (2005). *Thinking for a Living: How to Get Better Performance and Results From Knowledge Workers*. Boston, MA, USA: Harvard Business Press. Retrieved from [https://books.google.it/books?id=rpCEnGAGmRsC&dq=thinking+for+a+living&hl=p t-PT&source=gbs\\_navlinks\\_s](https://books.google.it/books?id=rpCEnGAGmRsC&dq=thinking+for+a+living&hl=p t-PT&source=gbs_navlinks_s)
- Docrated. (2016). Enterprise Search: The Benefits of Enterprise Search Software. Retrieved November 18, 2016, from <http://www.docrated.com/knowledge/enterprise-search-benefits-enterprise-search>
- Earley Information Science. (2016). Building the Business Case for Enterprise Search. Retrieved November 18, 2016, from <http://www.earley.com/knowledge/article/building-business-case-enterprise-search>
- European Commission. (2016a). *Data, Information and Knowledge Management at the European Commission. C(2016) 6626 final*. Brussels, Belgium.
- European Commission. (2016b). *Data, Information and Knowledge Management at the European Commission. Rolling Plan 2016 - 2017. SWD(2016) 333 final*.
- Feldman, S., Duhl, J., Marobella, J.R., & Crawford, A. (2005). *The hidden costs of information work*. Retrieved from <http://www.scribd.com/doc/6138369/Whitepaper-IDC-Hidden-Costs-0405>
- Formtek. (2016). Enterprise Search: Despite Benefits, Few Organizations Use Enterprise Search. Retrieved November 18, 2016, from <http://formtek.com/blog/enterprise-search-despite-benefits-few-organizations-use-enterprise-search/>
- Gartner. (2015). *Magic Quadrant for Enterprise Search*. Retrieved from <http://www.project-consult.de/files/Gartner Magic Quadrant for Enterprise Search Q3 2015.pdf>
- Google. (2016). Topical Search Engines. Retrieved December 7, 2016, from <https://developers.google.com/custom-search/docs/topical>
- Grand View Research. (2016). Enterprise Search Market Size To Reach \$8.90 Billion By 2024. Retrieved November 18, 2016, from <https://www.grandviewresearch.com/press-release/global-enterprise-search->

market

- Hewlett Packard Enterprise. (2016). *HPE IDOL for Search and Knowledge Discovery*. Retrieved from <https://www.hpe.com/h20195/v2/GetPDF.aspx/4AA6-3328ENW.pdf>
- Joint Research Centre. (2016). *JRC Strategy 2030*. Ispra.
- JRC Task Force on Migration. (2016). *Knowledge Centre on Migration and Demography (KCMD): Concept Note*.
- Liverpool University. (2016). Kritikos. Retrieved December 7, 2016, from <https://kritikos.liv.ac.uk/>
- Martin Butler Research. (2009). *The Business Value of Enterprise Search 2009*. Retrieved from <http://www.oracle.com/us/corporate/analystreports/infrastructure/059608.pdf>
- O'Dell, C., & Grayson, C. J. (1998). If only we knew what we know: identification and transfer of internal best practices. *California Management Review*, 40(3), 154–174.
- Peter F. Drucker. (1999). Knowledge-Worker Productivity: The Biggest Challenge. *California Management Review*. Retrieved from [http://www.forschungsnetzwerk.at/downloadpub/knowledge\\_workers\\_the\\_biggest\\_challenge.pdf](http://www.forschungsnetzwerk.at/downloadpub/knowledge_workers_the_biggest_challenge.pdf)
- Sassone, P. G. (1988). Cost benefit analysis of information systems: A survey of methodologies. In ACM Press (Ed.), *In Processings of the ACM SIGOIS and IEEECS TC-OA 1988 Conference on Office Information Systems* (pp. 126–133). New York.
- Search Explained. (2015). The Five C's of Enterprise Search. Retrieved November 18, 2016, from <http://searchexplained.com/the-five-cs-of-enterprise-search/#more-6431>
- Search Technologies. (2015). *Enterprise Search Fundamentals*.
- Search Technologies. (2016). Enterprise Search Benefits. Retrieved November 18, 2016, from <http://www.searchtechnologies.com/enterprise-search-benefits>
- Seddon, P.B., Staples, S., Patnayakuni, R., & Bowtell, M. (1999). Dimensions of information systems success. *Communications of the Association for Information Systems*, (2).
- Solinger, C., & Schubmehl, D. (2016). *Agile Information Discovery at AstraZeneca*. Retrieved from <http://www.sinequa.com/download/sinequa/use-case-astrazeneca-idc-11-2016.pdf>
- TATA Consultancy Services. (2015). *A Next Generation Search System for Today's Digital Enterprises*. Retrieved from <http://www.tcs.com/SiteCollectionDocuments/White-Papers/Next-Generation-Search-Digital-Enterprise-0815-1.pdf>
- White, M., & Nikolov, S. G. (2013). *Enterprise Search in the European Union: A Techno-economic Analysis*. Seville (Spain). Retrieved from <http://ftp.jrc.es/EURdoc/JRC78202.pdf>
- Wu, M., Turpin, A., Thom, J. A., Scholer, F., & Wilkinson, R. (2014). Cost and benefit estimation of experts' mediation in an enterprise search. *Journal of the Association for Information Science and Technology*, 65(1), 146–163. <http://doi.org/10.1002/asi.22951>

## List of abbreviations and definitions

COTS	Commercial off-the-shelf
EC	European Commission
EMM	European Media Monitor News Explorer
ES system	Enterprise Search Systems
EU	European Union
IT	Information Technologies
JRC	Joint Research Centre
KB	Knowledge Browser
KCMD	Knowledge Centre on Migration and Demography
KMD	Knowledge on Migration and Demography
OSS	Open Source Software
ROI	Return On Investment
TCO	Total Cost of Ownership
UIA platforms	Unified Information Access platforms



## List of figures

Figure 1: The “Gather and share” process .....	9
Figure 2: The “Retrieve” process .....	10
Figure 3: Architecture of the proposal .....	21
Figure 4: Costs. Conservative perspective scenarios. ....	27
Figure 5: Costs. Optimistic perspective scenarios. ....	27





## List of tables

Table 1: Structural elements .....	23
Table 2: Effort and benefits breakdown. No development. Conservative approach. ....	24
Table 3: Cost breakdown. No development. Conservative approach.....	24
Table 4: Effort and benefits breakdown. No development. Optimistic approach. ....	24
Table 5: Cost breakdown. No development. Optimistic approach. ....	25
Table 6: Effort and benefits breakdown. Development. Conservative approach. ....	25
Table 7: Cost breakdown. Development. Conservative approach. ....	25
Table 8: Effort and benefits breakdown. Development. Optimistic approach.....	26
Table 9: Cost breakdown. Development. Optimistic approach. ....	26



***Europe Direct is a service to help you find answers  
to your questions about the European Union.***

**Freephone number (\*):**

**00 800 6 7 8 9 10 11**

(\*) The information given is free, as are most calls (though some operators, phone boxes or hotels may charge you).

More information on the European Union is available on the internet (<http://europa.eu>).

## **HOW TO OBTAIN EU PUBLICATIONS**

### **Free publications:**

- one copy:  
via EU Bookshop (<http://bookshop.europa.eu>);
- more than one copy or posters/maps:  
from the European Union's representations ([http://ec.europa.eu/represent\\_en.htm](http://ec.europa.eu/represent_en.htm));  
from the delegations in non-EU countries ([http://eeas.europa.eu/delegations/index\\_en.htm](http://eeas.europa.eu/delegations/index_en.htm));  
by contacting the Europe Direct service ([http://europa.eu/europedirect/index\\_en.htm](http://europa.eu/europedirect/index_en.htm)) or  
calling 00 800 6 7 8 9 10 11 (freephone number from anywhere in the EU) (\*).

(\*) The information given is free, as are most calls (though some operators, phone boxes or hotels may charge you).

### **Priced publications:**

- via EU Bookshop (<http://bookshop.europa.eu>).

## JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.



**EU Science Hub**  
[ec.europa.eu/jrc](https://ec.europa.eu/jrc)



@EU\_ScienceHub



EU Science Hub - Joint Research Centre



Joint Research Centre



EU Science Hub



Publications Office

doi: 10.2788/33172

ISBN 978-92-79-64800-7